UNIVERSITÉ DE MONTRÉAL

CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

OPTIMAL STRATEGIES:
A NEW ASSIGNMENT MODEL FOR TRANSIT NETWORKS

by

Heinz Spiess and Michael Florian

Cette étude a été publiée grâce à une subvention du fonds F.C.A.R. pour l'aide et le soutien à la recherche.

Université de Montréal

Centre de recherche sur les transports - Publication #286
Département d'informatique et de
recherche opérationnelle - Publication #466

February 1983 - revised January 1988

Dépôt légal - 1er trimestre 1988

ABSTRACT

We describe a model for the transit assignment problem with a fixed set of transit lines. The traveler chooses the strategy that allows him to reach his destination at minimum expected cost. First we consider the case in which no congestion effects occur. For the special case in which the waiting time at a stop depends only on the combined frequency, the problem is formulated as a linear programming problem of a size that increases linearly with the network size. A label-setting algorithm is developed that solves the latter problem in polynomial time. Nonlinear cost extensions of the model are considered as well.

Key words : transit assignment; multiple-path; linear programming; polynomial algorithm.

- - - - - - - -

RÉSUMÉ

Nous décrivons un modèle pour le problème d'affectation sur un réseau de transport en commun qui comporte un ensemble fixe de lignes. Le voyageur choisit la stratégie qui minimise le coût requis pour atteindre sa destination. Nous envisageons d'abord le problème en l'absence de congestion. Dans le cas spécifique où le temps d'attente à un arrêt ne dépend que de la fréquence des lignes qui desservent cet arrêt, le problème est formulé comme un programme linéaire dont la taille croît linéairement avec celle du réseau. Un algorithme d'étiquetage, qui résoud ce problème en temps polynomial, est développé. Des extensions non linéaires du modèle sont également considérées.

Mots-clés : affectation transport en commun; chemins multiples; programmation linéaire; algorithme polynomial.

1. Introduction

The purpose of this paper is to propose a new model and develop a solution algorithm for the problem of determining route choice for transit riders in order to simulate flows over transit networks. First, we consider this problem when the relevant components of the travel time from origins to destinations are not subject to congestion effects. Whereas the corresponding problem for the car mode may be solved by an "all-or-nothing" assignment of the demand to the corresponding shortest paths, this is not a proper solution in the case of a transit network, due to the phenomenon of waiting for the arrival of a transit vehicle. This makes the transit route choice more complex. Also, in the case of private car, the computation of shortest paths is symmetric with respect to an origin/destination pair. This symmetry does not hold in the case of transit route choice.

The transit assignment problem has been studied by several authors in the past, either as a separate problem (e.g. see Dial, 1967; Chriqui, 1974; Rapp et al., 1976; Andreasson, 1976) or as a subproblem of more complex models, such as transit network design (see e.g Lampkin and Saalmans, 1967; Scheele, 1977; Mandl, 1979; Hasselström, 1981), or multimodal network equilibrium (Florian, 1977; Florian and Spiess, 1981). Computer programs, such as IGTDS (General Motors, 1978), TRANSCOM (Chapleau, 1974), TRANSEPT (Last and Leak, 1976), NOPTS (Rapp et al., 1976), the VOLVO system (Andreasson, 1976; Hasselström, 1981) and ULOAD (UMTA/FHWA, 1977), are each based on one of these methods and are widely used by transit companies and government agencies for planning purposes.

Most of the algorithms that have been proposed in the past may be classified as heuristic approaches to the problem, in the sense that the behavioral principle is stated qualitatively and leads directly to a computational procedure, where the model solved is either not stated or may never be stated formally. Often, these algorithms are variants of assignment procedures used for private car traffic on road networks (such as shortest path, stochastic multipath assignment), that are modified to reflect the waiting time phenomenon inherent to transit networks. In spite of the fact that for some of these algorithms the corresponding model that is solved may be identified, these are restricted version of the model proposed in this paper and therefore yield suboptimal solutions in relation with our model.

The subsequent sections of the paper are organized in the following way. In Section 2 we characterize a model for the transit traveler's route choice behavior which leads to the concept of "optimal strategy". This allows us to put into perspective and compare our

approach with some of the existing methods, highlighting the restrictions that are embedded in each of these algorithms. In Section 3 we state the mathematical programming model which corresponds to the optimal strategy and specialize it to the particular and tractable case when the waiting time at a boarding node depends only on the combined frequencies of the transit lines considered by the traveler. We show that the problem may be restated as a simpler linear program. The solution algorithm for this problem is presented in Section 4. It is shown that the algorithm may be applied directly to a network that represents the transit services and is of polynomial complexity, a property which makes it suitable for the solution of large scale problems. In Section 5 we consider nonlinear extensions of the model and Section 6 concludes the paper and outlines the potential for future research in related areas.

2. Strategies in Transit Route Choice

Let T denote a transit network consisting of a set of nodes, a set of transit lines, each defined as a sequence of nodes at which passengers may board and alight, and a set of walk links, each connecting two nodes. The times (or costs) that are associated with each walk link and each segment of a transit line are constant and known. At each node that is served by a line, the distribution of interarrival times of the vehicles of that line is known. Therefore, assuming that we also know the rate of arrival of the passengers at each node, we can construct the distribution of the waiting time for a vehicle of a given line. Hence, we can calculate the combined expected waiting time for the arrival of the first vehicle, among any given set of lines passing at the same node, as well as the probability of each line arriving first. The details of these calculations are given in detail in Section 3.

Note that each walk link may be replaced by a transit line of one link with a zero waiting time (very high frequency). In the following, the walk links will therefore not always be distinguished from the segments of the transit lines. In order to simplify the formulation, it is also assumed that the transit network T is strongly connected.

We now turn our attention to the transit traveler whose trip originates at node A and is destined to node B. We assume that he behaves in a way that minimizes his total expected travel time, which is the weighted sum of expected walk, wait, and in-vehicle time. This assumption seems natural and it also parallels the behavioral assumptions made for car traffic route choice. Since the weights may always be normalized to 1 by scaling the times used in

the definition of the network, we need not consider the weights separately. Thus, we will always use the term time, even if these "times" might actually include weight factors and should then more appropriately be called "generalized costs".

One could be tempted to reformulate the above problem as: "How does one find the path from node A to node B which minimizes the expected total travel time?", and by doing so one would already have changed the problem. We are accustomed to assume that the traveler selects a single path from a choice set which is the set of all possible paths. This is, of course, true in the case of a car driver, but this is not the case for a transit rider. He might, for instance, choose a set of paths and let the vehicle that arrives first determine which of the paths is actually used to get to his destination; in which case his choice does not correspond with the "a priori" specification of a path. His choice is therefore more complex than just the selection of a single path out of the set of possible paths. If the problem is limited to finding the path with the smallest expected travel time, a restriction of the problem is solved, yielding a suboptimal solution.

We call any element of the traveler's choice set a *strategy*. A strategy is a set of rules that, when applied, allows the traveler to reach his destination. The number and the kind of different strategies that the traveler may choose from depend on the information that is available to him during his trip. If no additional information becomes available during the trip, a strategy simply defines a path. For the transit network given in Figure 1, an example of such a strategy would be:

"Take line 2 to node Y; transfer to line 3 and exit at node B."

If the traveler, while waiting at a node, knows the line which is to be served next, a strategy may include subsets of lines and would then be:

"Take the next vehicle among lines 1 and 2; if line 1 was taken, exit at B; if line 2 was taken, transfer at node Y to any of the lines 3 or 4 and exit at B"

Of course, the more information available to the traveler (such as elapsed waiting time, arrival times of vehicles seen while waiting at a stop, information on other vehicles while riding in a vehicle by looking out of the window, ...), the more complex the strategies may become; for example:

"Wait up to five minutes for a vehicle of line 1; otherwise take line 2; if at node X you see a vehicle of line 3 (express bus), transfer to line 3, otherwise continue to

node Y and transfer there to any of the lines 3 or 4."

We make the assumption that the only information that is available to the traveler during his trip is that he finds out, while waiting at a node, which line is about to be served next; i.e., he may at that time decide whether or not to board this vehicle. Consequently, we only consider strategies that correspond to the second example given above. It is then possible to define a strategy of this type by specifying at every node of the network, except at the destination node, a non-empty set of attractive lines, and for each of these lines the (upstream) node at which the traveler alights. In the remainder of this paper we only consider strategies of this type. Note that when walk links are included in the network, a strategy may contain at a node an outgoing walk link instead of the set of attractive lines. Given a strategy, an actual trip is then carried out according to a mechanism that may be stated systematically as:

0: Set NODE to origin node.

1: Board vehicle that arrives first among the vehicles of the set of attractive lines at NODE.

2: Alight at the predetermined node.

3: If not yet at destination, set NODE to current node and return to step 1. Otherwise the trip is completed.

A strategy is said to be *feasible* if the graph defined by the attractive lines of the strategy does not contain cycles. Note that a strategy as defined above has a well determined destination node. The origin node, on the other hand, does not enter in the definition of a strategy. A strategy is therefore a set of rules that enables the traveler to reach his destination starting from any node in the network.

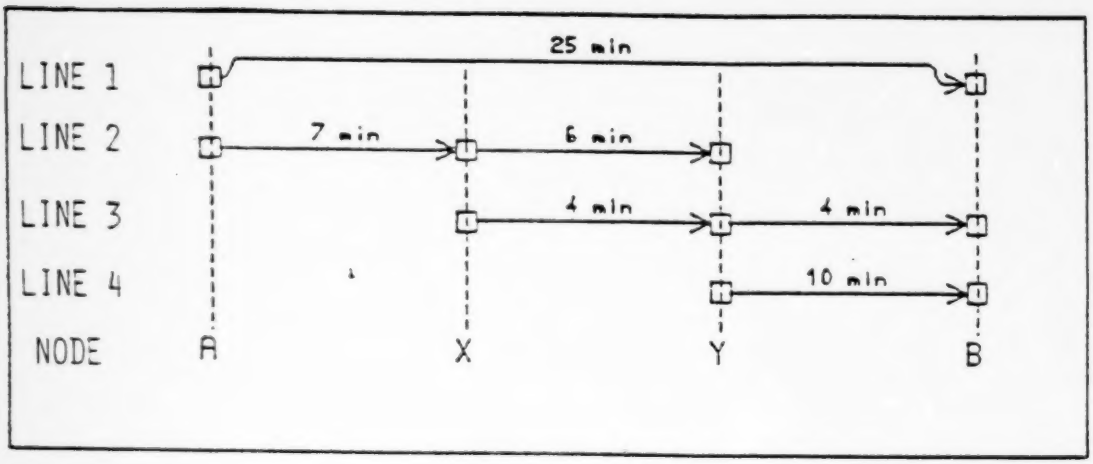

Figure 1: An example transit network.

Figure 1, which represents the transit lines used earlier to illustrate the concept of a strategy, shows a small transit network consisting only of four nodes and four transit lines. Since the network does not contain cycles, any strategy, as defined above, is feasible. There are 75 different strategies for reaching the destination node B from nodes A, X and Y. This is seen by the enumeration of all possible sets of attractive lines at each node except B, as shown in Table 1. However, there are only five distinct paths from A to B, four distinct paths from X to B and two distinct paths from Y to B.

In this example, the waiting time is computed by assuming that transit riders wait on the average half of the interarrival interval and that frequencies are combined linearly. The "line probability" is the probability that the line will be boarded and is the ratio of its frequency divided by the combined frequency.

Node	Attractive Lines (line → exit node)	Waiting Time (minutes)	Line Probabilities 1	2	3	4
A	1→B	6.0	1.00	-	-	-
A	2→X	6.0	-	1.00	-	-
A	2→Y	6.0	-	1.00	-	-
A	1→B, 2→X	3.0	0.50	0.50	-	-
A	1→B, 2→Y	3.0	0.50	0.50	-	-
X	2→Y	6.0	-	1.00	-	-
X	3→Y	15.0	-	-	1.00	-
X	3→B	15.0	-	-	1.00	-
X	2→Y, 3→Y	4.3	-	0.71	0.29	-
X	2→Y, 3→B	4.3	-	0.71	0.29	-
Y	3→B	15.0	-	-	1.00	-
Y	4→B	3.0	-	-	-	1.00
Y	3→B, 4→B	2.5	-	-	0.17	0.83

Table 1: All possible sets of attractive lines.

By using the in-vehicle times indicated in Figure 1 and the expected waiting time and line probabilities for each set of attractive lines at each node, as given in Table 1, we can now calculate the expected total travel time from A to B for each strategy. Figure 2 shows the optimal strategy, that is, the one minimizing the expected travel time. Note that the expected travel time is 27.75 minutes and that this strategy does not correspond to the choice of a single path. The volumes indicated on the line segments correspond to the assignment of one trip from A to B, and are denoted by v in Figures 2 to 4.

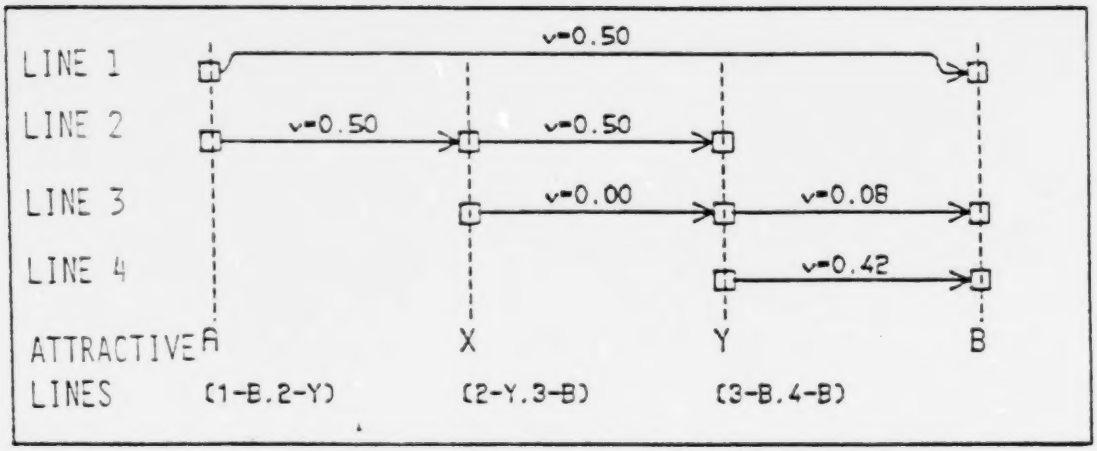

Figure 2: Optimal strategy. Expected travel time from A to B is 27.75 min.

If we compare this strategy with the one found by algorithms that are based on finding the least cost path (Dial, 1967; Fearnside and Draper, 1971), we see that the solution found by these methods, given in Figure 3, is clearly inferior to the optimal solution. The method described by Rapp et al. (1976) is also based on finding the shortest path, but the trips are then assigned to all efficient (in the case of the example all) paths according to Dial's multipath assignment method (Dial, 1971). Since the waiting times are those of individual lines (i.e. combined waiting times are not used), this method yields a solution with an expected travel time that is worse than the one of the least time path.

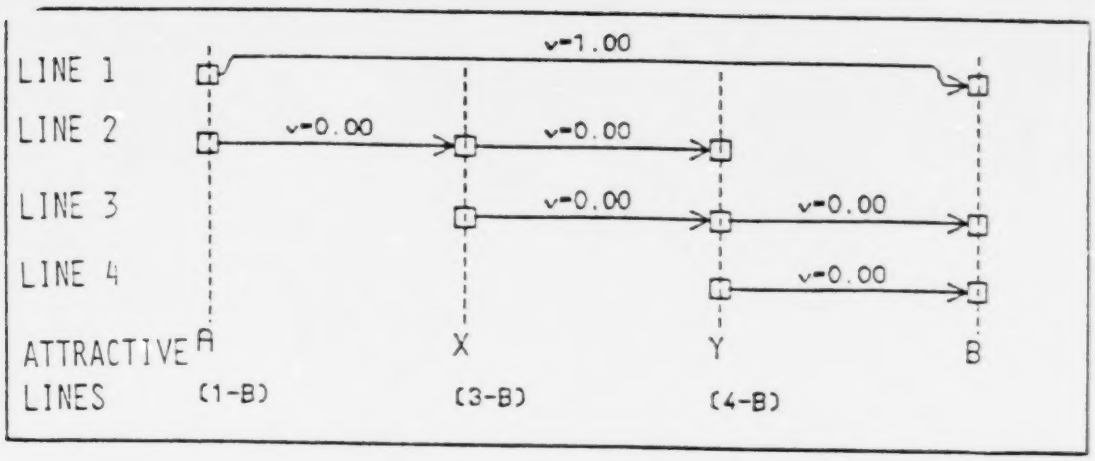

Figure 3: Best path to node B. Expected travel time A to B is 31 min.

Other methods described in the literature (Le Clercq, 1972; Chriqui, 1974) are also based on the idea of finding a least time path. But instead of defining the path in terms of consecutive segments of individual lines, paths are defined as a sequence of transfer nodes. Between two transfer nodes, the traveler's strategy may include more than one of the lines that connect the two transfer nodes directly. While in Le Clercq's method all direct lines are considered to be attractive, regardless of their in-vehicle time, Chriqui solves the problem of selecting the optimal subset of direct lines analytically (Chriqui and Robillard, 1975). The solution obtained by Chriqui's algorithm, which coincides for this example with Le Clercq's solution, is displayed in Figure 4. However, both of these models do not compute strategies of the type considered in this paper, which are more general.

Finally, the method proposed by Andreasson (1976) and later refined by Hasselström (1981), enumerates a certain number of distinct paths. Among these, the attractive paths are selected by a rather complicated set of heuristic rules that seem to work quite well in practice but do not guarantee optimality of the expected travel time. While all other methods mentioned above assign the demand from one origin to all destinations in one step, this method assigns the demand of each origin - destination pair separately.

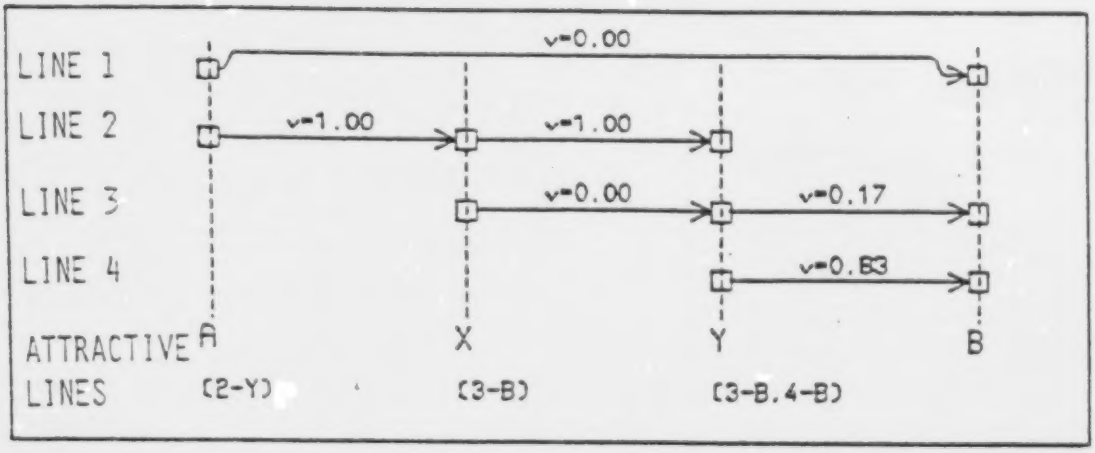

Figure 4: Best transfer sequence. Expected travel time A to B is 30.5 min.

In the sections that follow, a model is developed in which the traveler's choice set consists of all feasible strategies. It will become evident that, due to the asymmetry inherent to the transit trip, this model can only be solved efficiently by assigning the demands from all origins to a single destination at the same time.

3. The Transit Assignment Model

As we have seen in the preceding section, a transit trip consists in general of several trip components that may include some or all of the following:

- access from origin to transit stop,
- waiting for a vehicle,
- riding in a vehicle,
- alighting a vehicle,
- walking between two transit stops,
- egress from transit stop to destination.

A non-negative time (or cost) is usually used to quantify each of these trip components, with the exception of "waiting for a vehicle", which is quantified by using the statistical distribution of waiting times for the arrival of the first vehicle of a given transit line at a given stop.

In order to simplify the presentation, we consider in the following a generalized formulation of the problem, in which every trip component includes a constant time as well as a distribution of waiting times. The trip components are represented by links $a \in A$ of a network $G = (I, A)$ with nodes $i \in I$. Let A_i^+ (A_i^-) denote the set of outgoing (incoming) links at node $i \in I$. Each link $a \in A$ is characterized by the pair (c_a, G_a) where c_a is a non-negative link travel time and G_a the distribution function for the waiting time

$$G_a(x) = Prob\ \{waiting\ time\ on\ link\ a \leq x\ \}.$$

The functions $G_a(x)$ can be obtained from the distribution of interarrival times (headways) and the distribution of passenger arrival times. Let $h_a(x)$ be the density function of the distribution of interarrival times (headways) of the vehicles on link a. Let $H_a(x)$ denote its cumulative $H_a(x) = \int_0^x h_a(t)dt$. Assuming uniform arrival of passengers at the transit stop, the waiting time distribution for service on link a is $g_a(x) = (1-H_a(t)) \Big/ \int_0^\infty (1-H_a(t))dt$ with the cumulative $G_a(x) = \int_0^x g_a(t)dt$. For a link a that does not involve waiting we have

$$G_a(x) = \Delta(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x \geq 0. \end{cases}$$

Using this generalized formulation we can represent any transit network T. The mechanics that are involved in constructing the links and the nodes of the generalized network depend very much on the particularities of the transit network and the degree of aggregation considered.

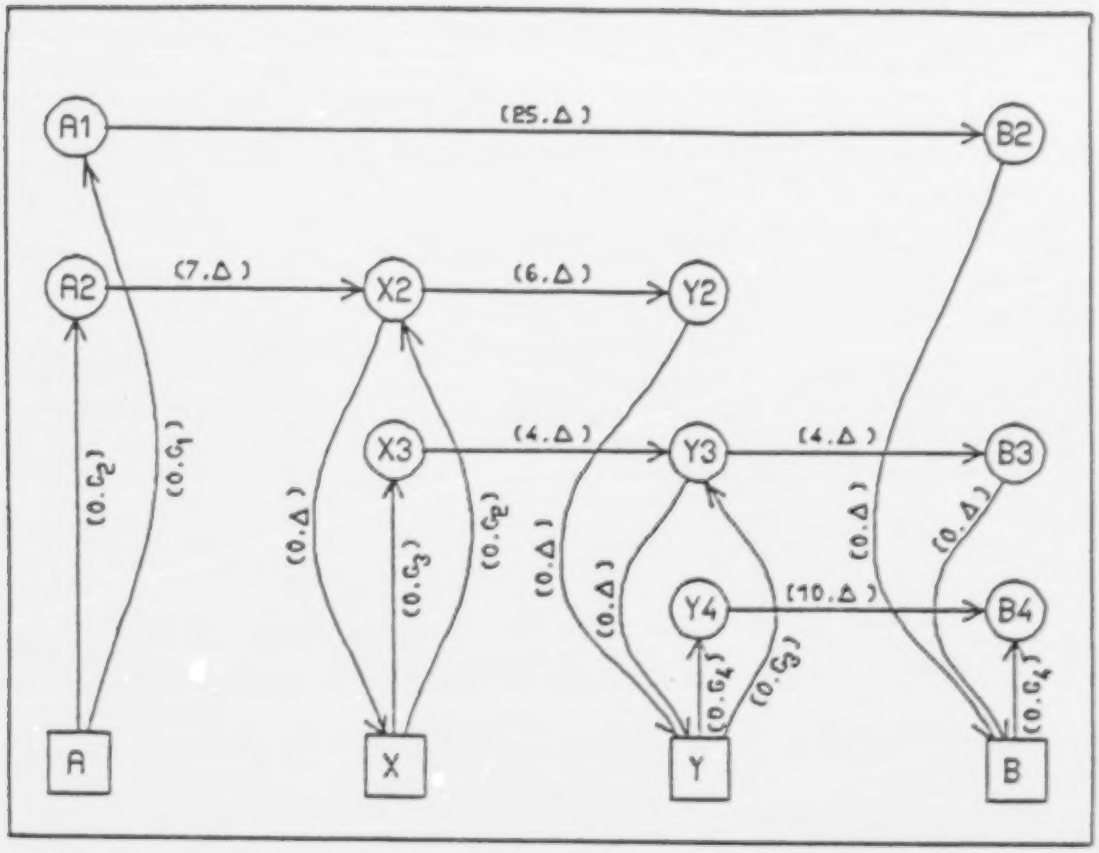

Figure 5: Link representation of example transit network.

Figure 5 shows such a network representation of the transit network used as an example in Section 2. In this case only three types of trip components are considered: waiting (no travel time), riding (no waiting) and alighting (no travel time and no waiting time). The links incident to and from nodes A, X, Y and B represent alighting and boarding respectively. By noting that a node which has only one incoming link a_1 with (c_{a_1}, G_{a_1}) and one outgoing link a_2 with $(c_{a_2}, G_{a_2} = \Delta)$ may be replaced with a single link a with $(c_a = c_{a_1} + c_{a_2}, G_a = G_{a_1})$, the network of Figure 5 may be reduced to the one shown in Figure 6.

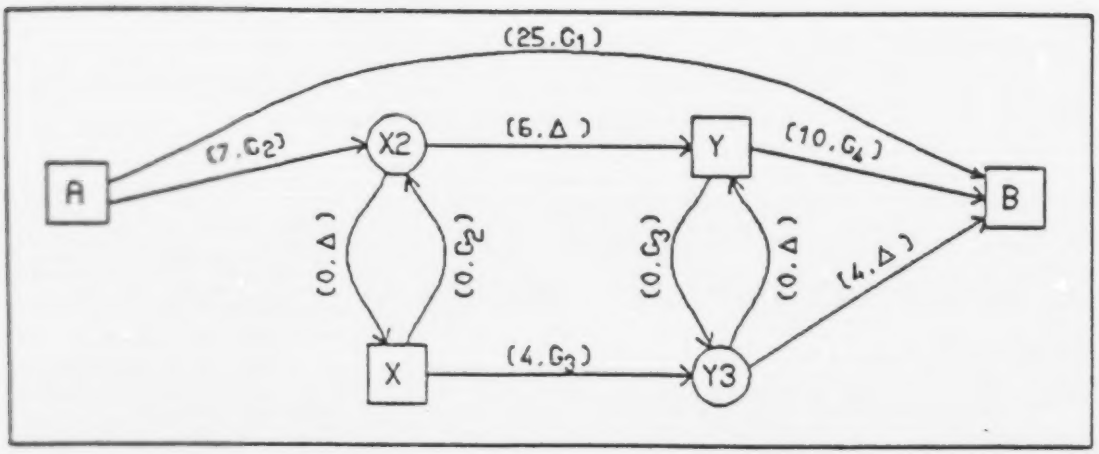

Figure 6: Simplified link representation of example transit network.

Since this generalized formulation no longer explicitly identifies transit lines but relies only on links, we can no longer use the terminology associated with the transit lines, such as waiting for a vehicle of a transit line. Instead, in the following, we will use the convention that each link is *served* by vehicles and that the passenger, instead of waiting for a vehicle of a certain line, waits for a *link to be served*.

For the transit route choice problem in this generalized form, a strategy to reach destination node r is defined by a partial network $G_r = (I, \bar{A})$ that contains only those links that will be used as a consequence of this strategy. We can therefore denote a strategy by the corresponding subset of links $\bar{A} \subseteq A$, or alternatively, by the partition of these links according to their i-node $\{\bar{A}_i^+ = A_i^+ \cap \bar{A}, i \in I\}$. Links not included in strategy $\bar{A}$ are never used. Among the links that are included in the strategy $\bar{A}$, at each node $i \in I$, a traveler boards the first vehicle that serves any of the links $a \in \bar{A}_i^+$. Hence, $\bar{A}_i^+$ corresponds to the set of attractive lines that was discussed in Section 2 and, of course $\bar{A}_i^+ \neq \emptyset$ for $i \neq r$. For a strategy to be *feasible* it must contain at least one path from each node to the destination node r and it must not contain any cycle.

Let $W(\bar{A}_i^+)$ denote the expected waiting time for the arrival of the first vehicle serving any of the links $a \in \bar{A}_i^+$. $W(\bar{A}_i^+)$ is called the *combined waiting time* of links $a \in \bar{A}_i^+$. Let further $P_a(\bar{A}_i^+)$ be the probability that link a is served first among the links $\bar{A}_i^+$. For convenience we define $P_a(\bar{A}_i^+) = 0$ for all $a \notin \bar{A}_i^+$. $W(\bar{A}_i^+)$ and $P_a(\bar{A}_i^+)$ depend on the distributions of waiting time $G_a(x)$ according to the following relationships:

$$W(\bar{A}_i^+) = \int_0^\infty \prod_{a \in \bar{A}_i^+} \{1 - G_a(x)\} dx$$

$$P_a(\bar{A}_i^+) = \int_0^\infty g_a(x) \prod_{\substack{a' \in \bar{A}_i^+ \\ a' \neq a}} \{1 - G_{a'}(x)\} dx.$$

The model that will be presented here is based on the assignment of the trips from all nodes toward a single node r, that we shall denote the *destination* node. This is different from most other assignment models, transit or highway, which always consider the trips from one origin node to all destination nodes. Let g_i, $i \in I - \{r\}$, be the demand (number of trips) from node i to the destination node r. In order to simplify the notation, we define

$$g_r = -\sum_{i \neq r} g_i .$$

We assume further that $g_i > 0$, $i \in I - \{r\}$, which prevents degeneracy of the variables to occur, thus allowing for more compact proofs in the analysis of the model. We note however, that all results presented below remain valid for $g_i \geq 0$, $i \in I - \{r\}$.

Given a strategy $\bar{A}$, the demand $g_i, i \in I$, is assigned to the network yielding link volumes $v_a, a \in A$. The volume at a node, which we denote $V_i, i \in I$, is the sum of the volumes of all incoming links and the demand at that node

$$V_i = \sum_{a \in A_i^-} v_a + g_i \ , \quad i \in I. \tag{1}$$

The node volume V_i is distributed on the outgoing links according to their link probabilities under strategy $\bar{A}$

$$v_a = P_a(\bar{A}_i^+)\, V_i \ , \quad a \in A_i^+ \ , i \in I. \tag{2}$$

Since $\sum_{a \in A_i^+} P_a(\bar{A}_i^+) = 1$, (1) and (2) ensure that the conservation of flow equation

$$\sum_{a \in A_i^+} v_a - \sum_{a \in A_i^-} v_a = g_i \quad , \quad i \in I \tag{3}$$

is satisfied.

The optimal strategy $\bar{A}^*$ is the strategy which minimizes the expected total travel time including waiting time. The most general formulation of the transit model is described by Spiess (1984).

In the following, we consider the special case in which the waiting time distribution of each link a (or transit line, in the original form of the problem) is quantified by a positive parameter f_a, which will be called the *frequency* of a link. The expected combined waiting time and the link probabilities are derived from the frequencies in the following way:

$$W(\bar{A}_i^+) = \frac{\alpha}{\sum_{a \in \bar{A}_i^+} f_a} \quad , \quad \alpha > 0 \, , \tag{4}$$

$$P_a(\bar{A}_i^+) = \frac{f_a}{\sum_{a' \in \bar{A}_i^+} f_{a'}} \quad , \quad a \in \bar{A}_i^+ . \tag{5}$$

The case α=1 corresponds to an exponential distribution of interarrival times of the vehicles with mean $1/f_a$ and a uniform passenger arrival rate at the nodes.

The case α=1/2 is an approximation of a constant interarrival time $1/f_a$ for the vehicles on link a. This measure of waiting time is the most widely used approach in practice (Chapleau, 1974; Dial, 1967; Le Clercq, 1972; Rapp et al., 1976; UMTA/FHWA, 1977), in spite of the fact that it is based on a rough approximation.

The factor α may, of course, also model the effect of different perceptions of waiting time and in-vehicle time. Since the link probabilities (5) are independent of the units in which f_a is specified, it is always possible to scale the frequencies by a factor $1/\alpha$. Without loss of generality we can therefore assume $\alpha = 1$ *in the following sections.*

In order to take advantage of the special case (4) and (5) for the statement of the model, we need to express the strategy $\bar{A}$ in terms of the 0-1 variables x_a

$$x_a = \begin{cases} 0 \ , \text{if } a \notin \bar{A} \\ \\ 1 \ , \text{if } a \in \bar{A} \end{cases} \quad , \quad a \in A. \tag{6}$$

We can now state the problem of finding the optimal strategy $\bar{A}^*$ as follows:

$$Min \quad \sum_{a \in A} c_a v_a + \sum_{i \in I} \frac{V_i}{\sum_{a \in A_i^+} f_a x_a} \tag{7}$$

subject to

$$v_a = \frac{x_a f_a}{\sum_{a' \in A_i^+} f_{a'} x_{a'}} V_i \quad , \quad a \in A_i^+ , i \in I, \tag{8}$$

$$V_i = \sum_{a \in A_i^-} v_a + g_i \quad , \quad i \in I, \tag{9}$$

$$V_i \geq 0 \quad , \quad i \in I, \tag{10}$$

$$x_a = 0 \text{ or } 1 \quad , \quad a \in A. \tag{11}$$

Let P1 be the problem defined by (7) to (11).

The number of variables and constraints are only of the order of the number of links and nodes of the network. Problem P1, at first sight, has a nonlinear objective function that is subject to nonlinear constraints, and the variables are partly continuous, partly integer. However, we shall demonstrate that this problem may be reduced to a much simpler linear programming problem by observing the following.

First, the non-negativity constraints for the node volumes (10) may be replaced by non-negativity constraints for the link volumes because of (9)

$$v_a \geq 0 \quad , \quad a \in A. \tag{12}$$

Next, by summing the constraints (8) for $a \in A_i^+$ we obtain

$$\sum_{a \in A_i^+} v_a = V_i \quad , \quad i \in I. \tag{13}$$

which may be used to substitute V_i in (9) to yield the conservation of flow constraints (3).

Then, by introducing new variables ω_i that denote the total waiting time for all trips at node i

$$\omega_i = \frac{V_i}{\sum_{a \in A_i^+} f_a x_a} \quad , \quad i \in I, \tag{14}$$

the variables V_i, $i \in I$, may be substituted.

The arguments leading to the justification of (12), (13) and (14) demonstrate the following.

Proposition 1: Problem P1 is equivalent to problem P2, which is

$$Min \quad \sum_{a \in A} c_a v_a + \sum_{i \in I} \omega_i \tag{15}$$

subject to (3), (11), (12) and

$$v_a = x_a f_a \omega_i \quad , \quad a \in A_i^+, i \in I. \tag{16}$$

As we can see, the objective function is now linear and the node volumes V_i, $i \in I$, no longer appear explicitly in the formulation. The 0-1 variables x_a are only used in constraints (16), which are the only nonlinear constraints of problem P2. These constraints may be relaxed by replacing (16) with

$$v_a \leq f_a \omega_i \quad , \quad a \in A_i^+, i \in I, \tag{17}$$

to yield the linear programming problem P3, which is

$$Min \quad \sum_{a \in A} c_a v_a + \sum_{i \in I} \omega_i \tag{18}$$

subject to

$$\sum_{a \in A_i^+} v_a - \sum_{a \in A_i^-} v_a = g_i \quad , \quad i \in I \tag{3}$$

$$v_a \leq f_a \omega_i \quad , \quad a \in A_i^+, i \in I, \tag{17}$$

$$v_a \geq 0 \quad , \quad a \in A. \tag{12}$$

Proposition 2: Problem P3 is equivalent to problem P2.

Proof: Problem P3 has $2|A| + |I| - 1$ linearly independent constraints and $|A| + |I| - 1$ variables, since ω_r may be arbitrarily set to zero. Each extreme point of the polytope defined by (3),(12),(17) has **at most** $|A|$ inequality constraints which are satisfied as strict inequalities. There are exactly two inequality constraints associated with each link, one from (12) and one from (17), that may not be satisfied simultaneously. It follows that **at least** one inequality constraint is satisfied as strict inequality for each link a. Thus, for each link a, $a \in A$, exactly one of the constraints from (12) and (17) is satisfied as a strict inequality, and it follows that the other one is satisfied as equality. This proves that each extreme point of P3 is feasible for P2. Since P3 is a relaxation of P2 this proves that the two problems are equivalent. □

Having finally cast the model as a linear programming problem, we can now examine its dual, problem P4, which is

$$Max \quad \sum_{i \in I} g_i u_i \tag{19}$$

subject to

$$u_j + c_a + \mu_a \geq u_i \quad , \quad a = (i,j) \in A, \tag{20}$$

$$\sum_{a \in A_i^+} f_a \mu_a = 1 \quad , \quad i \in I, \tag{21}$$

$$\mu_a \geq 0 \quad , \quad a \in A. \tag{22}$$

Let (v^*, ω^*) and (u^*, μ^*) denote any optimal solution of problems P3 and P4 which satisfy the weak complementary slackness conditions

$$(v_a^* - f_a \omega_i^*)\mu_a^* = 0 \quad , \quad a \in A_i^+, i \in I, \tag{23}$$

and

$$(u_j^* + c_a + \mu_a^* - u_i^*)v_a^* = 0 \quad , \quad a=(i,j) \in A. \tag{24}$$

So far we have only considered the case in which all links of the network involve waiting, that is, they have a finite frequency f_a. The case in which a subset of the links $\underline{A} \subseteq A$ involves no waiting may be modeled in two ways. A simple approach is to set $f_a = M$, where M is a sufficiently large number, so that the resulting waiting time $1/M$ can be ignored. It is

also possible to modify problems P2, P3 and P4, similarly to allow for links without waiting component. In problems P2 and P3 this results in dropping the constraints (16) and (17) for $a \in \underline{A}$ and adding explicitly non-negativity constraints for ω_i, $i \in I$. In the dual problem P4, constraint (20) applies only for links $a \notin \underline{A}$ and is replaced by $u_j + c_a \geq u_i$ for $a \in \underline{A}$. Constraint (21) becomes

$$\sum_{a \in (A_i^+ - \underline{A}_i^+)} f_a \mu_a \leq 1 \quad , \quad i \in I.$$

The complementary slackness conditions (23) and (24) reduce to $(u_j^* + c_a - u_i^*)v_a^* = 0$ for $a \in \underline{A}$.

Finally, it is worth while noting that, both in its primal and dual formulations, the transit assignment model presented here has a close resemblance to the shortest path assignment problem. In fact, the latter corresponds to a special case of the transit assignment problem for which none of the links of the network involves waiting, that is, $\underline{A} = A$. It is therefore not surprising that, as we shall see next, there is also a resemblance between the solution algorithms of the two problems.

4. The Solution Algorithm

In this section we present an algorithm that solves the transit assignment problem P4. The algorithm is composed of two parts. In a first pass, from the destination node to all origins, the optimal strategy $\overline{A}^*$ and the expected total travel times u_i^* from each node $i \in I$ to the destination node r are computed. In a second pass, from all origins to the destination, the demand is assigned to the network according to the optimal strategy.

The algorithm is stated below:

Part 1: Find Optimal Strategy

1.1 (Initialization) $u_i := \infty, i \in I - \{r\}$; $u_r := 0$;
$f_i := 0, i \in I$;
$S := A$; $\bar{A} := \emptyset$.

1.2 (Get next link) If $S = \emptyset$ then STOP,
else find $a=(i,j) \in S$ which satisfies
$u_j + c_a \leq u_{j'} + c_{a'}$, $a'=(i', j') \in S$;
$S := S - \{a\}$.

1.3 (Update node label) If $u_i \geq u_j + c_a$ then
$$u_i := \frac{f_i u_i + f_a(u_j + c_a)}{f_i + f_a},$$
$f_i := f_i + f_a$, $\bar{A} := \bar{A} + \{a\}$;
go to step 1.2.

Part 2: Assign Demand According to Optimal Strategy

2.1 (Initialization) $V_i := g_i, i \in I$;

2.2 (Loading) Do for every link $a \in A$, in decreasing order of $(u_j + c_a)$:
if $a \in \bar{A}$ then $v_a := \frac{f_a}{f_i} V_i$,
$V_j := V_j + v_a$,
else $v_a := 0$.

In step 1.1 the node labels u_i, i.e. the expected travel times to reach the destination, are set to infinity for all nodes except the destination, for which it is initialized to zero. The auxiliary variables f_i, $i \in I$, that contain the combined frequencies of all selected links at node i, are initialized to zero. The set S is used to identify the links that have not yet been examined, and the set $\bar{A}$ is used to identify the optimal strategy.

In step 1.2 the link nearest to the destination is selected among the links not yet examined. The time considered is $u_j + c_a$, that is, the time from the i-node of the link to the destination not including the waiting time at node i. If this time is smaller than the current time associated with node i, u_i, link a is included in the optimal strategy and u_i and f_i are

updated according to the formulas given in step 1.3. Note that when the label of a node i, u_i, is improved for the first time, we have $f_i u_i = 0 \cdot \infty$. In order to keep the algorithm as compact as possible we adapt the convention $0 \cdot \infty = 1$ whenever this case occurs. (When $\alpha \neq 1$ is used in (4), this convention becomes $0 \cdot \infty = \alpha$). Part 1 of the algorithm terminates when all links have been examined.

In the second part of the algorithm the demand from node i to the destination, g_i, is assigned to the network according to the strategy $\bar{A}$. This is done by assigning to each link $a \in \bar{A}$ the proportion of the node volume V_i that corresponds to its frequency. Since the links are processed in reverse topological order (decreasing $u_j + c_a$), the node volumes may be updated in parallel, hence making it possible to examine every link only once.

In order to prove that the solution $(\hat{u}, \hat{v})$ obtained by this algorithm is an optimal solution of problem P4, we proceed to show that $\hat{u}$ is a feasible solution of the dual problem P4, $\hat{v}$ is a feasible solution of the primal problem P3 and that $(\hat{u}, \hat{v})$ satisfy the complementary slackness conditions for problems P3 and P4. Optimality then results directly from the theorem of complementary slackness in linear programming.

Proposition 3: $\hat{u}$ is a feasible solution of the (dual) problem P4.

Proof: Let u_i^a and u_j^a be the values of u_i and u_j after step 1.3 has been performed for link $a=(i, j)$. Then

$$\begin{aligned} u_i^a &> u_j^a + c_a \quad \text{, if } a \in \bar{A}, \\ u_i^a &\leq u_j^a + c_a \quad \text{, else.} \end{aligned} \tag{25}$$

As the links are examined in increasing order of $u_j + c_a$ and $c_a \geq 0$, it follows that u_j will not be modified any more, therefore $u_j^a = \hat{u}_j$, $a \in A$. This is not true for u_i which may still be modified in step 1.3. Since u_i can only decrease, we have $\hat{u}_i \leq u_i^a$, $a \in A$. On the other hand, $\hat{u}_i = u_i^{a'}$, where $a' \in \bar{A}_i^+$ is the link included last into $\bar{A}$, which leads to

$$\hat{u}_i > u_j^{a'} + c_{a'} \geq u_j^a + c_a \quad , \quad a \in \bar{A}_i^+. \tag{26}$$

Thus, it follows that $\hat{u}$ satisfies

$$\begin{aligned} \hat{u}_i &> \hat{u}_a + c_a \quad \text{, if } a \in \bar{A}, \\ \hat{u}_i &\leq \hat{u}_j + c_a \quad \text{, else.} \end{aligned} \tag{27}$$

Furthermore, since we have shown that u_j will not change after it has been used in step 1.3, $\hat{u}_i$ may be expressed in terms of $\hat{u}_j$

$$\hat{u}_i = \frac{\sum_{a \in \bar{A}_i^+} f_a(\hat{u}_j + c_a) + 1}{\sum_{a \in \bar{A}_i^+} f_a} \quad , \quad i \in I. \tag{28}$$

In (28) we easily recognize the first term as the weighted average of the times for each link and the second term as the waiting time at node i. The dual variable $\hat{\mu}$ may be constructed as

$$\hat{\mu}_a = \begin{cases} \hat{u}_i - \hat{u}_j - c_a & , \text{ if } a \in \bar{A}, \\ 0 & , \text{ else.} \end{cases} \tag{29}$$

We can rewrite (28) as

$$\sum_{a \in \bar{A}_i^+} f_a(\hat{u}_i - \hat{u}_j - c_a) = 1 \quad , \quad i \in I. \tag{30}$$

Now, by inserting (29) into (30), it follows that

$$\sum_{a \in A_i^+} f_a \hat{\mu}_a = 1 \quad , \quad i \in I. \tag{31}$$

Since by construction $(\hat{u}, \hat{\mu})$ satisfies (20) and (22), it follows that this is a feasible solution of problem P4. □

Proposition 4: $\hat{v}$ is a feasible solution of the (primal) problem P3.

Proof: Let $\hat{f}_i$, $i \in I$, be the combined frequencies obtained by part 1 of the algorithm. According to step 1.3 they satisfy

$$\hat{f}_i = \sum_{a \in \bar{A}_i^+} f_a \quad , \quad i \in I. \tag{32}$$

Let the variables $\hat{\omega}_i$, $i \in I$, be defined as

$$\hat{\omega}_i = \frac{\hat{V}_i}{\hat{f}_i} \quad , \quad i \in I. \tag{33}$$

We can now express $\hat{v}_a$ in terms of $\hat{\omega}_i$ by using (33) and step 2.2 of the algorithm

$$\hat{v}_a = \begin{cases} f_a \hat{\omega}_i & , \text{if } a \in \bar{A}, \\ 0 & , \text{ else.} \end{cases} \tag{34}$$

Since the link volumes satisfy the conservation of flow equations and are non-negative by construction, it follows that $(\hat{v}, \hat{\omega})$ is a feasible solution of problem P3. □

Proposition 5: $(\hat{u}, \hat{v})$ are optimal.

Proof: Links included in strategy $\bar{A}$ satisfy

$$\hat{v}_a - f_a \hat{\omega}_i = 0 \quad , \quad a=(i,j)\in \bar{A}, \tag{35}$$

and

$$\hat{u}_j + c_a + \hat{\mu}_a - \hat{u}_i = 0 \quad , \quad a=(i,j)\in \bar{A}. \tag{36}$$

where for links not included

$$\hat{v}_a = 0, \quad \hat{\mu}_a = 0 \quad , \quad a \notin \bar{A}. \tag{37}$$

Hence the solutions $(\hat{v}, \hat{\omega})$ and $(\hat{u}, \hat{\mu})$ satisfy the weak complementary slackness conditions (23) and (24) and are therefore optimal solutions of problems P3 and P4. Since $(\hat{v}, \hat{\omega})$ is an extreme point solution of P3, $(\hat{v}, \bar{A})$ is an optimal solution of problem P2 by proposition 2. □

The case, in which a subset of links, $\underline{A} \subseteq A$ does not involve waiting for service, may again be handled by either setting the frequencies f_a of these links to a very large number, or by adapting the algorithm to this special case. The latter approach yields two slight modifications of steps 1.3 and 2.2 of the algorithm. If $a\in\underline{A}$, the updating of the variables in step 1.3 becomes

$$u_i := u_j + c_a \ ,$$

$$f_i := \infty \ , \ \bar{A}_i^+ := \{a\}.$$

If a link $a\in\underline{A}$ is part of the optimal strategy, its volume is calculated as $v_a := V_i$ in step 2.2. No proof is given for this modified version of the algorithm, since it follows the same lines as the proofs for propositions 3, 4 and 5.

The following example shows how the algorithm is applied to the small network used as an example in Sections 2 and 3. Since the waiting times and the link probabilities given in Table 1 correspond to the special case (4) and (5), it is possible to assign frequencies f_a to the links. Figure 7 shows the network as well as the link times and frequencies (c_a, f_a).

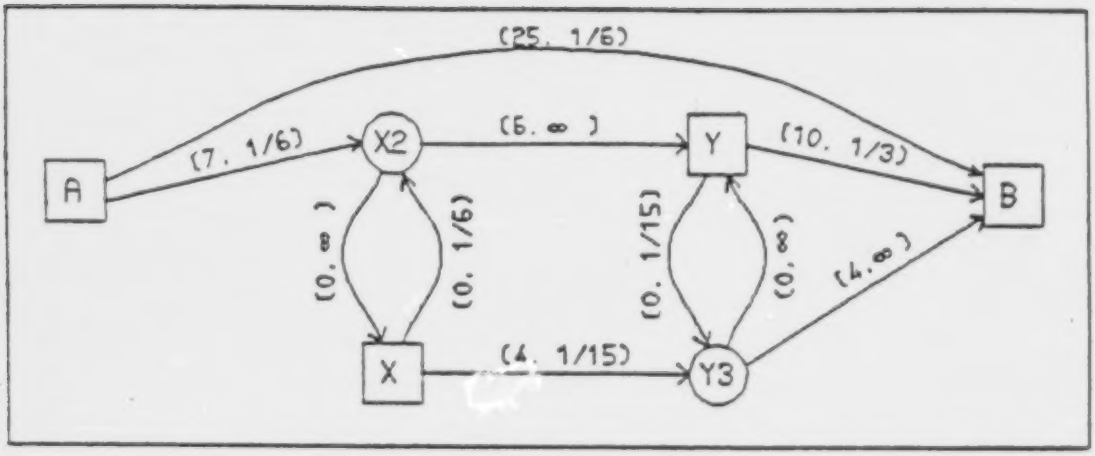

Figure 7: Transit network with link costs and frequencies (c_a, f_a).

Note that some links have no waiting component ($f_a = \infty$), hence the modified version of the algorithm is therefore used. Table 2 shows how the optimal strategy is computed (part 1 of the algorithm).

#	Node Labels (u_i, f_i)						Link with Min u_j+c_a			
	A	X2	X	Y3	Y	B	$a=(i,j)$	f_a	u_j+c_a	$a \in \bar{A}$
1	∞,0	∞,0	∞,0	∞,0	∞,0	0,0	(Y3,B)	∞	4.0	yes
2	"	"	"	4.0,∞	"	"	(Y,Y3)	1/15	4.0	yes
3	"	"	"	"	19.0,1/15	"	(X,Y3)	1/15	8.0	yes
4	"	"	23.0,1/15	"	"	"	(Y,B)	1/3	10.0	yes
5	"	"	"	"	11.5,2/5	"	(Y,Y3)	∞	11.5	no
6	"	"	"	"	"	"	(X2,Y)	∞	17.5	yes
7	"	17.5,∞	"	"	"	"	(X,X2)	1/6	17.5	yes
8	"	"	19.0,7/30	"	"	"	(X2,X)	∞	19.0	no
9	"	"	"	"	"	"	(A,X2)	1/6	24.5	yes
10	30.5,1/6	"	"	"	"	"	(A,B)	1/6	25.0	yes
	27.75,1/3	17.5,∞	19.0,7/30	4.0,∞	11.5,2/5	0,0				

Table 2: Find optimal strategy for example network.

In Table 3 the demand, one trip from A to B, is assigned to the network (part 2 of the algorithm).

#	Link (i,j)	Volume	Node Volumes A	X2	X	Y3	Y	B
			1.00	0.00	0.00	0.00	0.00	-1.00
10	(A,B)	0.50	1.00	0.00	0.00	0.00	0.00	-0.50
9	(A,X2)	0.50	1.00	0.50	0.00	0.00	0.00	-0.50
8	(X2,X)	0.00	1.00	0.50	0.00	0.00	0.00	-0.50
7	(X,X2)	0.00	1.00	0.50	0.00	0.00	0.00	-0.50
6	(X2,Y)	0.50	1.00	0.50	0.00	0.00	0.50	-0.50
5	(Y,Y3)	0.00	1.00	0.50	0.00	0.00	0.50	-0.50
4	(Y,B)	0.42	1.00	0.50	0.00	0.00	0.50	-0.08
3	(X,Y3)	0.00	1.00	0.50	0.00	0.00	0.50	-0.08
2	(Y,Y3)	0.08	1.00	0.50	0.00	0.08	0.50	-0.08
1	(Y3,B)	0.08	1.00	0.50	0.00	0.08	0.50	0.00

Table 3: Assign demand on example network.

The solution found by the algorithm corresponds to the optimal strategy displayed in Figure 2.

We examine next the complexity of the algorithm. Let n and m be the number of nodes and links in the network considered. The following table shows the complexity of every step of the algorithm and the number of times it is executed.

Step	Complexity	#Times
1.1	$O(n+m)$	1
1.2	$O(m)$	m
1.3	$O(1)$	m
2.1	$O(n)$	1
2.2	$O(m)$	1

Table 4: Complexity of the steps of the algorithm.

The complexity of the entire algorithm for one destination node is determined by step 1.2 and is $O(m^2)$. Note that in step 2.2 no additional computations are needed to establish the order in which the links are processed, since it is the inverse of the order used in part 1 of the algorithm. Step 1.2 can, of course, be implemented by using a heap structure, thus reducing its complexity to $O(log\ m)$ and the complexity of the algorithm to $O(m\ \log m)$ per destination node.

If the network contains more than one destination node, the algorithm is applied once for each destination. In this case working memory space can be saved by modifying the algorithm slightly so that the variables v_a, $a \in A$, are used directly to store the accumulated link volumes for all destinations as follows:

(I) Initialize link volumes $v_a := 0$, $a \in A$.

(II) Assign the demand to each destination using the algorithm above with step 2.2 modified in the following way:

Do for every link $a \in A$, in decreasing order of $(u_j + c_a)$:

$$\text{if } a \in \bar{A} \text{ then } v_a := v_a + \frac{f_a}{f_i} V_i,$$

$$V_j := V_j + \frac{f_a}{f_i} V_i.$$

The complexity of the algorithm for all destinations is $O(|R|\ m\ \log m)$, where $|R|$ denotes the number of destinations with non-zero demand.

5. Nonlinear Extensions of the Model

We now turn our attention to the version of the transit assignment problem in which the link travel times c_a are no longer constants, but are continuous functions $c_a(v)$, $a \in A$, of the vector of link flows v. Such functions may model queuing delays at stops, since passengers may not board the first arriving vehicle, the discomfort "cost" of passengers in crowded vehicles or the travel time of vehicles under different load conditions.

In the following we discuss briefly a general formulation of the nonlinear model and show that when $c_a(v) = c_a(v_a)$, $a \in A$, there exists an equivalent convex optimization

problem. Then we discuss the advantages and drawbacks of these formulations.

Let v_a^r denote the volume on link a, $a \in A$, that is associated with destination node r, $r \in R$. Since the total volume v_a on a link is the sum of the volumes v_a^r to all different destinations $r \in R$,

$$v_a = \sum_{r \in R} v_a^r \quad , \quad a \in A, \tag{38}$$

it is no longer possible to separate the problem by destination.

Let the demand from node i, $i \in I$, to destination r, $r \in R$, be denoted g_i^r and let K_r denote the set of all feasible strategies to reach destination r; h_i^k is the part of the demand g_i^r that is assigned to the network using strategy $k \in K_r$. Thus, we can state the conservation of flow equations

$$\sum_{k \in K_r} h_i^k = g_i^r \quad , \quad i \in I, \ k \in K_r, \ r \in R. \tag{39}$$

We assume, in the same way that we did for constant link times, that every traveler chooses the strategy that minimizes the expected travel time of his trip. However, in the nonlinear model, the travel times depend on the vector of link flows v and therefore on the strategies chosen by the other travelers.

A user optimal travelers' behavior can be characterized by equilibrium conditions paralleling those derived from Wardrop's second principle for the car traffic network equilibrium (Wardrop, 1952; Nguyen, 1974). Let s_i^k denote the expected travel time from node i to destination r using strategy $k \in K_r$ and let u_i^r denote the minimum expected travel time from i to r; the equilibrium conditions may now be stated as

$$\left.\begin{array}{l} s_i^{k^*} = u_i^{r^*}, \ \text{if} \ h_i^{k^*} > 0 \\ s_i^{k^*} \geq u_i^{r^*}, \ \text{if} \ h_i^{k^*} = 0 \end{array}\right\} \quad , \ i \in I, \ k \in K_r, \ r \in R. \tag{40}$$

It is rather straight forward to show that these equilibrium conditions may be restated in the form of a variational inequality (by following the methodology of Smith (1979)). However, this does not lead to an immediately applicable model since the mapping of $c(v)$ is not likely to be strictly monotone and, hence, there may exist multiple equilibria. In addition, the choice of a solution algorithm is not evident, since most solution algorithms for variational

inequality problems require the mapping $c(v)$ to satisfy conditions which are at least as strong as monotonicity in order to assure convergence. Thus, this version of the problem is still open for further research.

When the link travel times are continuous increasing functions $c_a(v_a)$ of the total volume on link a, there exists the following convex cost differentiable equivalent optimization problem P5

$$Min \sum_{a \in A} \int_0^{v_a} c_a(x)dx + \sum_{i \in I} \sum_{r \in R} \omega_i^r \tag{41}$$

subject to (38) and

$$v_a^r \leq f_a \omega_i^r \ , \quad a \in A_i^+, \ i \in I, \ r \in R, \tag{42}$$

$$\sum_{a \in A_i^+} v_a^r = \sum_{a \in A_i^-} v_a^r + g_i^r \ , \quad i \in I, \ r \in R, \tag{43}$$

$$v_a^r \geq 0 \ , \quad a \in A, \ r \in R. \tag{44}$$

We refer the interested reader to Spiess (1984) for the proof that the above problem indeed satisfies the equilibrium conditions (40). Since this is a convex minimization problem, it may be solved using any of the standard techniques available for this class of problems such as the linear approximation method (Frank and Wolfe, 1956) or its PARTAN variant (Florian, Guélat and Spiess, 1987).

In the following, we outline briefly a solution algorithm that is based on the linear approximation method. Since it does not require the explicit storage of the destination dependent volumes v_a^r, it is especially suited to handle large transit networks with many destination nodes. We note that the linear approximation of the objective function (41) results in a problem that is separable by destination. For each destination the resulting subproblem is equivalent to the problem with constant link times as discussed in Section 3, and may therefore be solved efficiently using the algorithm given in Section 4. The resulting algorithm to solve problem P5 may be outlined as follows:

0. (Initialization) Find a feasible solution (v^0, ω^0), where v^0 denotes a feasible vector of total flows v_a and the scalar ω^0 denotes the corresponding total waiting time $\sum_{i \in I} \sum_{r \in R} \omega_i^r$.

Set $k := 0$.

1. (Subproblem) Set $k := k + 1$.

Compute $(\hat{v}, \hat{\omega})$ by applying the algorithm of Section 4 with link times $c_a = c_a(v_a^{k-1})$ for each destination $r \in R$.

2. (Line Search) Find λ^k which minimizes the objective function of P5 on the line segment

$(1-\lambda)(v^{k-1}, \omega^{k-1}) + \lambda(\hat{v}, \hat{\omega})$, $0 \leq \lambda \leq 1$.

3. (Update) Set $(v^k, w^k) := (1-\lambda^k)(v^{k-1}, w^{k-1}) + \lambda^k(\hat{v}, \hat{\omega})$.

If $|| (v^k, w^k) - (v^{k-1}, w^{k-1}) || < \epsilon$ then STOP

else go to step 1.

It is worthwhile to note that problem P5 includes as a special case the network equilibrium problem. When no link is subject to waiting, the link frequencies are infinite and the expected waiting times are zero.

This approach allows us to model some of the consequences of transit vehicle capacity by using discomfort functions similar to the volume delay functions used in the equilibrium road assignment to model street capacities. However, there are some limitations to this approach that have to be kept in mind:

- In this model, all passengers aboard a transit vehicle suffer the same degree of discomfort. In reality the passenger that boards the vehicle first has a better chance of getting a seat and experiences therefore a different level of discomfort than a passenger that boards the vehicle when no more seating space is available.
- The waiting time is not directly affected by the transit volumes. Every passenger may board the first arriving vehicle and queuing delays at stops because of full vehicles are not considered by the model.
- The time the vehicle spends at the transit stop may depend in reality on the number of passengers that alight and board at that stop. This dependence is not considered either by this model.

6. Conclusions

The model and algorithm described in this paper have been implemented as part of the EMME/2 system, an interactive-graphic method for urban transportation planning, developed at the "Centre de recherche sur les transports" of the Université de Montréal (Babin et al., 1982; Spiess, 1984). The data structure has been adapted to avoid the explicit construction of the links and nodes of the generalized transit network (see Spiess, 1984). The program is written in FORTRAN and has been ported to the following computers: IBM PC/AT with or without Definicon DSI-32/DSI-780 co-processors, SUN-2 and SUN-3 workstations, Masscomp, HP9000-500, the entire line of VAX/VMS computers and IBM 309x mainframe computers.

The execution times vary from one computing environment to another, as indicated in Table 5, which gives the execution times for the transit assignment algorithms on models of the transit networks of Winnipeg, Man., Montreal, Qué., Stockholm Inner City, Sweden, Oakland, CA and Ottawa, Ont.

City	Computer	# Transit Lines	# Line Segments	# Zones	# Generalized Links	Execution Time (seconds)
Winnipeg	Microvax II	67	4,205	154	15,590	450.30
Winnipeg	DIS/780 on IBM/AT	67	4,205	154	15,590	186.70
Montreal	IBM 3091	283	18,720	700	73,082	712.58
Stockholm Inner City	SUN/3	38	1,058	185	4,509	149.96
Oakland (AC Transit)	Microvax II	108	6,452	272	22,074	1602.61
Ottawa (OC Transpo)	Vax 11/780	140	7,698	380	78,758	1713.15

Table 5: Execution times.

The EMME/2 system is used for transit planning by many agencies in the United States, Canada and Europe. Some of the applications have been documented (see Gault, 1986 and Sullivan, 1987). The results of the assignment have proven to be realistic, when compared to observed line flows and total number of boardings.

As each transit assignment calibration and validation study is likely to be different, depending on the data available for calibration and the services offered, we shall not discuss this issue in any detail in this paper.

The applicability of the transit model presented in this paper is most satisfying, since in addition to it being based on sound behavioral assumptions, which lead to a clear mathematical programming formulation, it is solved by a polynomial time algorithm which has time and memory requirements comparable to those of a shortest path algorithm. Evidently, it may be and has been applied to very large scale transit networks.

Since the model with flow dependent travel times includes as a special case the road network equilibrium assignment model, it unifies the formulation of transit and road assignment algorithms. It may also find application to other network assignment problems in which waiting for service has to be considered, such as freight shipment by truck, rail or plane.

The basic model presented in this paper has been adapted for modeling the waiting time for infrequent services, where passengers adjust their arrival time at the stop in order to reduce their waiting time and for coordinated transfer time nodes, where the waiting time is not determined by the line frequency - all without changing the basic model formulation and the solution algorithm. Thus, the model is quite flexible and may be applied to a variety of special situations that occur in the supply of transit services. Transfers from one transit service to another, which involves payment of an additional fare may be modeled as a walk link with a travel time that is the equivalent, in time units, of the additional fare paid.

This model may serve as a starting point for future research. Several extension of the model are possible and have the potential of enriching the models available for transit planning.

In Section 5 we mentioned several limitations of the nonlinear model that considers congestion effects. Most of the difficulties could be overcome by letting the link times c_a depend not only on the volume of link a, $c_a(v_a)$, but also on the volumes of the other links, $c_a(v)$. This model may be stated as a variational inequality. Unfortunately, a unique solution to the problem is only ensured if the mapping $c(v)$ is strictly monotone. Since queuing phenomena that we try to model lead to non-monotone mappings $c(v)$, multiple solutions may occur, rendering the problem difficult to solve and the solutions non-trivial to interpret. Another interesting approach to the modeling of congestion effects would model the

frequencies f_a to be decreasing functions of the total volume aboard the vehicle, $f_a(v)$. This "effective frequency" would take into account increased waiting times due to overcrowded vehicles that prevent boarding and force the passengers to wait for another vehicle. This extension of the model is not well understood yet.

Several authors have proposed transit network design models that aim to optimize the frequencies of the lines by stating a normative objective function (Fearnside and Draper, 1971; Last and Leak, 1976; Scheele, 1977). Replacing the simplistic route choice models that are used in these models by our transit assignment model could lead to more realistic transit network design models.

References

Andreasson I., *A Method for the Analysis of Transit Networks,* **2nd European Congress on Operations Research**, edited by Mark Roubens, North Holland, 1976.

Babin A., Florian M., James-Lefebvre L. and Spiess H., *EMME/2: Interactive Graphic Method for Road and Transit Planning,* **Transportation Research Record** 866, pp 1-9, 1982.

Chapleau R., *Réseaux de Transport en Commun: Structure Informatique et Affectation,* Ph.D. thesis, Dép. I.R.O., Université de Montréal, 1974.

Chriqui C., *Réseau de Transport en Commun: Les Problèmes de Cheminement et d'Accès,* Ph.D. thesis, Dép. I.R.O., Université de Montréal, 1974.

Chriqui C. and Robillard P., *Common Bus Lines,* **Transportation Science**, 9, pp 115-121, 1975.

Dial R.B., *Transit Pathfinder Algorithm,* **Highway Research Record**, 205, pp 67-85, 1967.

Dial R.B., *A Probabilistic Multipath Traffic Assignment Model which Obviates Path Enumeration,* **Transportation Research**, Vol 5, 1971.

Fearnside K. and Draper D.P., *Public Transport Assignment - A New Approach,* **Traffic Engineering and Control**, pp 298-299, November 1971.

Florian M., *A Traffic Equilibrium Model of Traffic by Car and Public Transit Modes,* **Transportation Research** 10, pp 166-179, 1977.

Florian M. and Spiess H., *On Two Mode Choice / Assignment Models,* Publication No. 177, Centre de Recherche sur les Transports, Université de Montréal, 1981.

Florian M., Guélat, J. and Spiess H., *An Efficient Implementation of the "PARTAN" variant of the Linear Approximation Method for the Network Equilibrium Problem*, **Networks** 17, pp 319-339, 1987.

Frank M. and Wolfe P., *An Algorithm for Quadratic Programming*, **Naval Research Logistics Quarterly** 3, pp 95-110, 1956.

Gault, H. *The Installation of Computerized Service Planning at OC Transpo*, Systems Planning Department Working Paper #3, Ottawa-Carleton Regional Transit Commission, November 1986.

General Motors Transportation System Center, *Interactive Graphic Transit Design System (IGTDS): User's Manual*, General Motors Technical Center, Warren, MI, 1978.

Hasselström D., *Public Transportation Planning - A Mathematical Programming Approach*, Ph.D. thesis, Department of Business Administration, University of Gothenburg, Sweden, 1981.

Lampkin W. and Saalmans P.D., *The Design of Routes, Service Frequencies and Schedules for a Municipal Bus Undertaking: A Case Study*, **Operations Research Quarterly**, Vol 18, No 4, pp 375-397, 1967.

Last A. and Leak S.E., *Transept: a Bus Model*, **Traffic Engineering and Control**, pp 14-20, January 1976.

Le Clercq F., *A Public Transport Assignment Method*, **Traffic Engineering and Control**, pp 91-96, June 1972.

Mandl C.E., *Evaluation and Optimization of Urban Public Transportation Networks*, presented at the **3rd European Congress on Operations Research**, Amsterdam, Netherlands, 1979.

Nguyen S., *An Algorithm for the Traffic Assignment Problem*, **Transportation Science**, 8, pp 203-216, 1974.

Rapp M.H., Mattenberger P., Piguet S. and Robert-Grandpierre A., *Interactive Graphic System for Transit Route Optimization*, **Transportation Research Record**, 619, 1976.

Scheele C.E., *A Mathematical Programming Algorithm for Optimal Bus Frequencies*, Institute of Technology, University of Linkoping, Sweden, 1977.

Smith M.J., *Existence, Uniqueness and Stability of Traffic Equilibria*, **Transportation Research B** 18, pp 295-304, 1979.

Spiess H., *Contributions a la théorie et aux outils de planification des réseaux de transport urbain*, Ph.D. thesis, Département d'Informatique et de Recherche Operationnelle, Université de Montréal, 1984

Sullivan, E.C., *Systems Analysis for the AC Transit Comprehensive Plan: Model Development and Validation,* Research Report UCB-ITS-RR-87-14, Institute for Transportation Studies, University of California, Berkeley, November 1987.

UMTA/FHWA, *UTPS Reference Manual,* U.S. Department of Transportation, 1977.

Wardrop J.G., *Some Theoretical Aspects of Road Traffic Research,* **Proc. Inst. Civil Engineers, Part II** , 1, pp 325-378, 1952.